Usos documentales del marcado de texto periodístico: NEWSML y NITF

Tony Hernández
Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid
Madrid, 126
28903 Getafe (Madrid)
Tel: 91 624 92 52 Fax: 91 624 92 12
tony@bib.uc3m.es

David Rodríguez
Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid
Madrid, 126
28903 Getafe (Madrid)
Tel: 91 624 92 51 Fax: 91 624 92 12
pirio@bib.uc3m.es

Resumen

Se expone una breve descripción de los lenguajes de marcado de contenidos periodísticos NITF y NewsML, incluyendo una breve evolución histórica de los lenguajes previos que conducen a la creación de ambos, una descripción de su estructura básica, una relación de las semejanzas y diferencias entre dichos lenguajes, una muestra de sus características básicas, entre las que destacan la modularidad, la flexibilidad y la gran capacidad de descripción documental, y una reflexión sobre las implicaciones que puede tener su uso para los documentalistas de prensa, entre las que destaca la tarea, aún pendiente, de normalización de los términos y vocabularios empleados.

Palabras clave

Periodismo, lenguajes de marcado, NITF, NewsML, normalización, documentalistas de prensa.

1. Introducción

Desde la aparición de XML en 1998, el uso de este metalenguaje como herramienta documental se ha multiplicado. El hecho de contar con la experiencia de su hermano mayor SGML, pero con una menor complejidad, ha permitido que surjan numerosos lenguajes de marcado de texto aplicados a las más diversas disciplinas, o bien que ciertos lenguajes basados en SGML sean convertidos a XML.

El ámbito del periodismo, o mejor, de las telecomunicaciones aplicadas al periodismo, es uno de los sectores en los que el marcado de texto tiene una mayor antigüedad. No obstante, si bien los primeros lenguajes de este tipo datan de comienzos de los 70, la aplicación de metalenguajes es mucho más reciente.

En particular, la apuesta que tiene visos de ser mayoritaria en la comunidad de la prensa digital en línea parece ser el tándem formado por NewsML y, en menor medida, por NITF. NewsML está llamado a ser un estándar, al menos, en su uso por parte de tres de las mayores agencias de prensa del mundo: la británico-estadounidense Reuters (su impulsora), la francesa AFP y la alemana DPA. Otras agencias e instituciones, como la estadounidense AP y la cadena japonesa de periódicos Mainichi Newspapers, ya trabajan con NewsML.

En la presente comunicación, se expondrá, en primer lugar, por qué los lenguajes de marcado periodístico constituyen el pasado, y el futuro, de la comunicación periodística a través de medios digitales en línea. Posteriormente, se muestran, en síntesis, los fundamentos de NITF y de NewsML: cuál es su uso concreto, su estructura básica y las características que relacionan –no necesaria, pero sí convenientemente– y distinguen a ambos, así como sus ventajas: la modularidad, que permite varios niveles de descripción documental; y la capacidad de integrar múltiples documentos en un solo contenedor. Por último, se discuten las implicaciones, desde los puntos de vista periodístico y documental, del uso de ambos lenguajes, así como algunas reflexiones, a modo de conclusión, sobre el papel que puede jugar los documentalistas en el uso y aplicación de estos lenguajes.

2. El uso de los marcados de texto con fines periodísticos

Ya a comienzos de la década de los 70 del pasado siglo, surgieron, con los primeros terminales en las redacciones de medios informativos, los intentos por transmitir la información directamente desde los ordenadores de las agencias de prensa hasta los ordenadores del medio correspondiente (ya fuera de prensa, radio o televisión). Esta pretensión no era baladí: por ejemplo, el volumen de noticias que puede publicar cualquier periódico del mundo, basándose en informaciones de agencia, puede oscilar entre el 25%, en las secciones de grandes periódicos cubiertas por periodistas del medio, y el 100%, en secciones donde el periódico no cuenta con corresponsales o colaboradores. Este último caso corresponde a secciones, como Internacional o Economía, en periódicos locales de todo el mundo.

Para conseguir la comunicación electrónica directa agencia-medio, surgen lenguajes que meramente hacían hincapié en la transmisión de los contenidos entre ordenadores, y

no en la descripción documental de dichos contenidos. Lenguajes como ANPA-1170 (1974) o ANPA-1228 (1976), convertido en IPTC TEC-7901 (1979), y sus variaciones, incluían ya metadatos –aunque aún no los llamaran así– especialmente enfocados a detalles sobre el tamaño del fichero, el modo de transmisión, el comienzo y fin de la misma, los saltos de línea, etc. En las versiones más avanzadas, comenzaron a surgir algunos metadatos temáticos.

Fue a comienzos de los 90 cuando surgió el interés de los medios informativos, y sobre todo, las agencias, por aplicar la transmisión de contenidos a algo más que a mero texto. Así surgen las primeras clasificaciones temáticas que, bajo el nombre de *Subject Codes,* comienza a promover el IPTC, uno de los organismos internacionales que más ha apostado por la aplicación de las telecomunicaciones al periodismo, junto con la estadounidense NAA y la asociación internacional IFRA, que reúnen a editores de periódicos en Estados Unidos y en todo el mundo.

Pero el factor que determinó la entrada definitiva del marcado de texto para la producción de contenido periodístico digital fue la explosión de Internet como objeto de consumo masivo. HTML se quedó pronto pequeño para quienes querían emplear lenguajes de marcado más potentes, con un mejor marcado estructural. Así, surgió en 1995 NIUTF, elaborado por el IPTC, basado en SGML y con una fuerte influencia de HTML.

El último escalón tecnológico, por el momento, fue la aparición de XML en 1998. El subconjunto de SGML logró que, en muchos campos, lenguajes basados en SGML se reconvirtieran a versiones bajo XML. Así, NIUTF pasó a ser NITF, en 1999.

No obstante, NITF seguía siendo un lenguaje, como se muestra a continuación, enfocado solo hacia el texto. Para fomentar el etiquetado, descripción y transmisión de contenidos en cualquier formato, IPTC creó NewsML en 2000.

A continuación, se muestran algunos rasgos de ambos lenguajes.

3. NewsML

NewsML es un lenguaje de contenedores de noticias digitales. Es decir, con NewsML no pueden crearse noticias en un formato concreto: es preciso disponer de ellas previamente en otros formatos.

¿Cuál es entonces su utilidad? Transportar paquetes de contenidos periodísticos, sea cual sea su formato y su forma de difusión: texto, imagen y sonido. Por ejemplo, si una agencia quiere ofrecer un paquete informativo sobre la final del Campeonato Mundial de Fútbol, formado por tres versiones textuales de una crónica en tres idiomas diferentes, más algunas fotos, más dos vídeos con imágenes de los goles marcados, más las opiniones de los protagonistas en ficheros de audio... todo eso puede ir en un único contenedor; pero cada una de las piezas (cada texto, cada vídeo, cada fichero de audio) debe ser creado en su propio formato. Siguiendo con el ejemplo, podría incluir un texto en NITF, su versión en PDF, etc.; una imagen en Real Player, otra en Windows Media...

La utilidad de NewsML, pues, es la misma, aunque con mayor complejidad, que la de los primeros lenguajes de marcado periodístico: la transmisión de contenidos periodísti-

cos entre medios, con la salvedad de que, ahora, esos contenidos son mucho más complejos, y también lo es el lenguaje utilizado.

Más aún, NewsML, como vemos en su estructura básica, permite contener más de un paquete informativo (o, simplificando, más de una noticia) en un solo documento. La figura 1 puede ser un ejemplo simplificado de NewsML:

```
<NewsML>
    <NewsItem>
        <NewsComponent >
            <ContentItem>
                (Aquí puede el contenido de una pieza informativa, en
                cualquier formato, o una referencia a ese contenido,
                que esté físicamente en otro fichero.)
            </ContentItem>
        </NewsComponent >
    </NewsItem >
</NewsML>
```

Figura 1. Estructura básica de un documento NewsML

Básicamente, se entiende:

- Que cada pieza está incluida en un elemento ContentItem.
- Que el conjunto de varios ContentItem (un texto, una foto, un vídeo y un gráfico) pueden formar parte de un News Component (una noticia).
- Que varias noticias pueden estar contenidas en un NewsItem.

Un documento NewsML puede llevar además toda una serie de metadatos que, en resumen, pueden ser de tres tipos:

- Datos relativos a la transmisión del documento NewsML en conjunto: quién lo envía, a quién, cual es su prioridad, la fecha de caducidad...
- Datos sobre el documento en cuestión, o sobre partes determinadas: por ejemplo, en cada una de las piezas que se incluye, pueden añadirse elementos que describan el tema, o los protagonistas, o las relaciones de esa pieza con otras en el mismo documento...
- Datos sobre cómo se normalizan los datos incluidos en otros elementos de NewsML. Por ejemplo, si en un NewsItem se incluye una noticia sobre fútbol, se puede incluir un elemento, o Topic, que describa temáticamente el deporte sobre el que trata la noticia. Ese elemento puede incluir un código; será entonces, un elemento añadido, o Catalog, el que indique cual es la clasificación o vocabulario del que forma parte el código, y dónde se encuentra la clasificación completa, en Internet.

La mayor parte de los metadatos que forman parte de NewsML pueden situarse en múltiples lugares del documento, aplicados a una sola pieza o a varias. Como puede deducirse, NewsML tiene dos características básicas:

- Una estructura modular, que además permite que las piezas estén situadas físicamente dentro del documento NewsML o fuera de él, unidas mediante referencias a objetos externos.
- La posibilidad de realizar una descripción estructural, aunque sea solo para distinguir cada una de las piezas (pero no sus partes), y una descripción semántica, también de tipo general.

4. NITF

NITF (News Industry Text Format) es un lenguaje para la descripción de contenidos periodísticos en modo texto. NITF permite la descripción a dos niveles de profundidad: de modo general, mediante metadatos añadidos al margen del texto, en un elemento *head* con sus correspondientes elementos hijos; y dentro del propio texto, incluido en el elemento *body,* englobando al texto correspondiente mediante un elemento concreto. Por ejemplo, si queremos normalizar dos referencias distintas a una misma persona, puede procederse como en los siguientes ejemplos:

```
El presidente <person name.given= "George Walker" name.family="Bush">
Bush</person>...

<person name.given= "George Walker" name.family="Bush">George
Bush</person>...
```

Figura 2. Dos formas de uso del elemento *person* en NITF

De este modo, además, no puede confundirse la referencia a esa persona con la referencia a otra persona con un nombre similar (en el ejemplo, se evitaría la confusión con Herbert George Bush, padre del anterior, y también Presidente de los Estados Unidos entre 1988 y 1992).

La estructura básica de un documento NITF se muestra en la siguiente figura:

```
<nitf>
      <head>
            <!– Aquí irían metadatos de carácter general: fecha del documento,
            fecha de caducidad, datos sobre el emisor de la noticia, derechos de
            uso y reproducción, palabras clave... Todos ellos son opcionales. –>
      </head>
```

```
<body>

  <body.head>
        <!– Aquí podrían ir datos sobre el documento en general: titular
        principal, autoría, fecha, resumen... –>
  </body.head>

        <!– El contenido "real" del documento está en el siguiente elemento –>

  <body.content>

        <!– El documento puede llevar uno o más bloques. Cada bloque es una
        pieza informativa. –>

        <block>
              <hl2>Gran partido del Almendralejo...</hl2>
              ...
        </block>
        <block>
              <hl2>Datos estadísticos del partido...</hl2>
              ...
        </block>
        <block>
              <hl2>Entrevista al goleador, Paquito...</hl2>
              ...
        </block>
  </body.content>

</body>

</nitf>
```

Figura 3. Estructura básica de un documento NITF

5. Relación entre NITF y NewsML

El formato NITF es el recomendado, pero no necesario, para insertar piezas de texto en un documento NewsML. Ello se debe a que ambos comparten una misma forma de estructurar los documentos, a pesar de tener usos muy diferentes, como se ha visto:

- Ambos permiten tanto la descripción semántica como la estructural (de modo más acentuado, en NITF).
- Ambos permiten el uso de metadatos dentro del propio documento o en un anexo aparte.

- Ambos poseen una gran modularidad en la inserción de metadatos, incluso dentro del propio documento.
- Ambos pueden contener más de una pieza informativa.

6. Subject Codes

Tanto NITF como NewsML permiten la descripción temática mediante códigos, a partir de múltiples clasificaciones. El propio IPTC ha creado una clasificación, que denomina *Subject Codes,* basada en tres niveles de descripción:

- El primer nivel indica si se trata de una noticia textual, de datos (por ejemplo, una clasificación deportiva) o un contenido aún no publicable.
- El segundo nivel indica el tipo de contenido, desde un punto de vista periodístico, en el que se incluye la pieza: noticia, análisis, biografía (perfil), opinión...
- El tercer nivel incluye varios grados de clasificación, tanto numéricos como alfabéticos: descripción del tema a tres niveles de profundidad, abreviatura y código numérico. Por ejemplo:

Categoría principal (Subject Code)	**Código númerico (Subject Reference Matter)**	**Tema (Subject Matter Name)**	**Abreviatura**
Arte, Cultura y Ocio	01005000	Cine	CIN

Figura 4. Un ejemplo de *Subject Code*.

No obstante, este modelo de clasificación es muy mejorable:

- Las categorías principales, 16, no abarcan todos los ámbitos del saber humano (teniendo en cuenta que el periodismo, por definición, es enciclopédico), aunque sí los más habituales en los medios informativos.
- Los temas no están aún completamente desarrollados. Solo algunas categorías, como Deportes, se han desglosado en profundidad y, aún así, manifiestan un profundo sesgo anglófilo y, más concretamente, estadounidense.

7. Implicaciones de NITF y de NewsML para los documentalistas

El uso de NITF y NewsML supone una serie de implicaciones tanto desde un punto de vista periodístico como desde una perspectiva meramente documental. Sin intención de agotar, por razones de espacio, los interrogantes que pueden plantearse, señalamos a continuación algunos de ellos:

- ¿Cómo y cuando se realizará la inserción de metainformación que puede añadirse tanto en los elementos dentro de los contenidos NewsML y NITF, como en los metadatos añadidos? ¿Quién realizará esa inserción?

- ¿Qué grado de profundidad tendrá el marcado de texto, especialmente en NITF?
- ¿Cómo se definirán los vocabularios empleados para la descripción de los metadatos que pueden ser normalizados?
- ¿Cuál es la validez de los Subject Codes?
- ¿Cómo afectará el uso de ambos lenguajes a los archivos de los medios electrónico, y a sus capacidades de recuperación de información?
- ¿Cuál será el papel que jugarán los documentalistas de prensa en relación con estas tecnologías, vistos los interrogantes anteriores?

Tratamos de ofrecer algunas respuestas a estas cuestiones, a continuación.

El primer problema que se plantea es compaginar la gran capacidad documental de ambos lenguajes, especialmente de NITF, con la cada vez mayor velocidad de producción de noticias en línea. Resulta obvio que, al menos, el marcado interno de las noticias no puede ser realizado manualmente de modo simultáneo a su producción.

Los documentalistas, pues, tendrán que hacer hincapié más en la definición de los mecanismos automáticos de marcado (en colaboración con técnicos como los informáticos o ingenieros de telecomunicaciones, y usuarios como los periodistas o los gestores y administrativos del medio), haciendo hincapié en las ventajas que puede presentar un marcado documental cada vez más preciso.

En la misma línea se enmarca la segunda de las implicaciones planteadas, el grado de profundidad del marcado. NITF y NewsML pueden llegar a una profundidad de marcado extrema, señalando prácticamente cualquier posible término de búsqueda que aparezca en un documento, con su correspondiente descripción contextual y su respectiva normalización. Ello plantea dudas sobre cual es el grado de ruido y de silencio que puede causar la mayor o menor abundancia de descripción documental tanto en la recuperación mediante motores de búsqueda como en la generación y uso de hiperenlaces navegables, teniendo en cuenta que, a medio plazo, estos enlaces no serán únicamente unidireccionales, como en HTML, sino multidireccionales, mediante la aplicación de técnicas como XLink.

Especialmente, la citada normalización dependerá del modo en el que se generen y utilicen los vocabularios y clasificaciones correspondientes, ya sean de uso exclusivo del medio periodístico, ya sean clasificaciones de tipo más general, como los citados Subject Codes de IPTC. Éstos últimos, no obstante, deben ser necesariamente complementados, tanto por su parcialidad temática como por el hecho de permanecer incompletos en su desarrollo. La comunidad periodística, y especialmente, los documentalistas en este campo, deben plantear la posibilidad de establecer una clasificación lo más universal posible para permitir el intercambio de información; la utilidad de este tipo de clasificaciones es bien conocida en biblioteconomía.

Los documentalistas deben aprovechar el vacío aún existente para establecer dicha clasificación, adquiriendo el debido protagonismo que, en ciertos ambientes, especialmente tecnológicos, se les ha querido robar, hasta cuestionar en ocasiones su propia existencia.

Por último, la combinación de todos los factores indicados (definición de correctas estrategias de marcado; generación de hiperenlaces útiles y graduables en dificultad, adaptados a cada usuario; correcta normalización de las descripciones profundas del con-

tenidos; capacidad para establecer mecanismos de intercambio de información; protagonismo en estas tareas de los documentalistas) debe repercutir en la mejor implantación de herramientas de recuperación de información y, especialmente, de lectura y navegación para los diversos usuarios de la información periodística: no solo el tradicional receptor de medios periodísticos, sino también, otros usuarios no menos importantes. Uno de ellos, más tradicional: los miembros de la redacción del medio, que podrán interrelacionar mejor los contenidos pasados, presentes y futuros, para generar productos informativamente más complejos. Y otros usuarios, que solo en los últimos años han comenzado a ser considerados en la bibliografía sobre el tema: los gestores del medio, que podrán extraer todo tipo de datos estadísticos, y aplicar diversos parámetros de evaluación cuantitativa sobre el rendimiento del medio, a partir de los datos digitales almacenados.

El trabajo en el que se basa esta comunicación ha sido parcialmente financiado por el proyecto TEL1999-0207 de la CICYT.

Bibliografía consultada

ALLEN, D.; Mohr, W. (1998). Considerations for the Semantic Markup with the NITF, 1/2/1998. http://www.darmstadt.gmd.de/topas/publications/Allen_Moehr_1998.pdf [Consulta: 30/09/2002].

ALLDAY, T. (2001). "NewsML. Enabling a standards-led revolution in news publishing?" Technical Review, nº 287, 1-8.

ARNOULD, V. (2002) "Y lo mejor está por llegar". Técnicas de prensa, nº 17, 22-23.

ARUNDALE, J.; Withey, R. (2002). "Los sofisticados archivos permiten a los usuarios buscar mejor contenido". Técnicas de prensa, nº 14, págs. 14-17.

COLE, D.M. (1999). "Twenty years in the making". TechNews, vol. 5, nº 3, http://www.naa.org/technews/TNArtPage.cfm?AID=2864 [Consulta: 25/06/2002]

FIELD, J.L. (2001) "IT Standards in the News. The history of international IT standards for news". http://www.iptc.org/site/history.html [Consulta: 08/09/2002].

FOURNIER, V. (2002). "NewsML, un estándar multimedia para la distribución de información". Técnicas de prensa, nº 17, págs. 16-20.

IPTC (2001). News Industry Text Format, http://www.nitf.org, 2002 [Consulta: 09/09/2002].

IPTC (2002). NewsML in action. http://www.newsml.org, 1/2/2001 [Consulta: 09/09/2002].

ITCH, Takashi (2002). "NewsML, palabra de moda en JANPS". Técnicas de prensa, nº 17, 30.

SHIPSIDE, S. (2000). "Los nuevos sistemas redefinen a los modernos documentalistas". Técnicas de prensa, nº 4, 8-11.

QUINN, S. (2002). "Prepararse para los cambios de las redacciones del mañana". En Técnicas de prensa, nº 17, págs. 34-35.

REUTERS (2002): NewsML Showcase, http://about.reuters.com/newsml/ [Consulta: 28/09/2002].